title

Tziporah Horowitz

Johns Hopkins University

Methods

There are several ways to measure the association between a risk factor and the binary outcome of contracting a disease. The following sections discuss the four approaches used in this analysis to determine the associated risk of various factors in the context of patients contracting heart disease. Sections 1–3 refer to table 1 for simplification.

	Diseased	Healthy
Exposed	D_E	H_E
Unexposed	D_U	H_U

Table 1. Contingency Matrix

1. Risk Difference

Often considered the simplest approach for measuring associated risk, *risk difference* or *absolute risk difference* (ARD) is the difference in the outcome rates between patients with the risk factor and patients without the risk factor (Telke & Eberly, 2011). Using the matrix in 1, risk difference can be defined mathematically as:

$$\text{ARD} = \frac{D_E}{D_E + H_E} - \frac{D_U}{D_U + H_U}$$

While the risk difference is easy to compute, its interpretation is often misleading and can only explain the associated risk between a single factor and the target.

2. Relative Risk

Similar to risk difference, *relative risk* compares the outcome rates between patients with the risk factor and patients without the risk factor. However, relative risk is computed as a ratio (RR) rather than a difference (Telke & Eberly, 2011). The risk ratio is defined as:

$$\text{RR} = \frac{D_E/(D_E + H_E)}{D_U/(D_U + H_U)}$$

Relative risk is a useful statistic because it quantifies the probability of a patient with exposure contracting the disease relative to a patient without exposure. Risk ratios that are close to 1 indicate that the risk of contracting the disease for an exposed patient is the same as the risk for an unexposed patient. In contrast, risk ratios that are far from 1 indicate that there is an association between the variables. This allows one to create a confidence interval using the hypothesis test,

$$H_0 : RR = 1$$

$$H_1 : RR \neq 1$$

The risk ratio is considered a valid measure of relative risk for studies in which the sampling is dependent on the exposure of interest such as, randomized controlled trials or cohort and cross-sectional studies (Gallis & Turner, 2019). Like risk difference, relative risk can only explain the associated risk between a single factor and the target.

3. Odds Ratio

Often confused with risk ratio, *odds ratio* compares the statistical odds of the outcome in the exposed group to that of the outcome of the unexposed group. It is defined mathematically as:

$$\text{OR} = \frac{D_E/H_E}{D_U/H_U}$$

Like the risk ratio, odds ratios that are close to 1 indicate no association between exposure and contracting the disease, and odds ratios that are far from 1 indicate that there is an association between the variables. One can also create a confidence interval for the odds ratio using a similar hypothesis test to that of the risk ratio, such that

$$H_0 : OR = 1$$

$$H_1 : OR \neq 1$$

While the odds ratio is typically considered the "only valid measure of relative association in traditional case-control studies" (Gallis & Turner, 2019), it is frequently misinterpreted as the risk ratio. However, in cases where the risk factor is relatively small ($< 10\%$), the odds ratio approximates the risk ratio:

$$\frac{D_E/(D_E + H_E)}{D_U/(D_U + H_U)} \approx \frac{D_E/H_E}{D_U/H_U} \quad \text{if } D_E \text{ and } D_U \text{ are small}$$

The odds ratio can be applied in multi-parameter settings when computed in a logistic regression analysis, due to its inherent calculation of the logit (or log-odds) function. To obtain the odds ratio of a logistic regression model, one simply has to exponentiate the coefficients.

4. Marginal Effects

A *marginal effect* (or incremental effect) is the change in the probability that an outcome occurs as the risk factor changes by one unit. It is often used in logistic regression analysis and other GLM's to explain the incremental risk associated with each factor (Norton, Dowd, & Maciejewski, 2019). Marginal effects are determined by taking the partial derivative of the regression equation with respect to each variable. They are simpler to interpret than odds ratios and are easier to compare across different studies. There are many ways to represent the marginal effect for a sample, the most common of which is the *average marginal effect* across all patients in the dataset.

The Analysis

For this analysis, the *Heart Disease Dataset* was collected from kaggle.com (Lapp, 2019). The dataset includes data that was compiled from four databases in 1988 and consists of 1025 rows and 14 columns: 13 predictors and 1 target. The predictors include 5 continuous variables: age, resting blood pressure, serum cholestoral (in mg/dl), maximum heart rate achieved, and ST depression induced by exercise relative to rest (oldpeak); and 8 categorical variables: sex, chest pain type, fasting blood sugar $>$ 120 mg/dl (true or false), resting electrocardiographic results, exercise induced angina (yes or no), the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy, and thal (normal, fixed defect, reversible defect). Figure 1 shows the *skimpy* summary of all 14 variables and figure 2 shows the distributions of predictor variables when compared to the target, a binary indicator for the patient having heart disease.

Prior to computing any of the above measures of associated risk, the data was imported into Python (3.9) using `pandas` and dummy variables were created for the categorical columns. Individual risk differences and risk ratios were then computed for each predictor variable. The dataset was then divided into a $3:2$ train-test split so that a logistic regression model can be fit. Two models were fit using the training set to the objective function,

$$p = \boldsymbol{\beta X}$$

via `statsmodels`. One was computed with the odds ratio, and the other with marginal effects. The models were then tested with the remainder of the data and scored using `sklearn`.

Results

	Risk Difference	Risk Ratio	RR 95% Confidence Interval
age	-0.2892	0.5671	(0.5008, 0.6423)
sex	-0.3036	0.5809	(0.5204, 0.6484)
chest pain $\{1, 2\}$	0.4813	2.597	(2.2719, 2.9685)
resting blood pressure > 130	-0.0888	0.839	(0.7411, 0.9499)
serum cholestoral > 250 ml/dl	-0.1512	0.7374	(0.6473, 0.8401)
fasting blood sugar > 120 mg/dl	-0.0577	0.8893	(0.7415, 1.0666)
resting electrocardiographic results $\{0, 1\}$	0.3178	2.5891	(0.9394, 7.1362)
maximum heart rate achieved > 150	0.4067	2.364	(2.0403, 2.7391)
exercise induced angina	-0.4633	0.3076	(0.2483, 0.3809)
oldpeak	-0.3773	0.435	(0.3708, 0.5103)
slope	0.1453	1.384	(1.027, 1.8649)
major vessels colored $\{0, 1, 2\}$	0.2593	1.94	(1.3731, 2.741)
thal $\{2, 3\}$	0.1882	1.5567	(1.118, 2.1675)

Dep. Variable:	target	**No. Observations:**	615
Model:	Logit	**Df Residuals:**	602
Method:	MLE	**Df Model:**	12
Date:	Thu, 24 Nov 2022	**Pseudo R-squ.:**	0.5058
Time:	15:10:58	**Log-Likelihood:**	-210.48
converged:	True	**LL-Null:**	-425.93
Covariance Type:	nonrobust	**LLR p-value:**	1.073e-84

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
age	0.0272	0.014	1.983	0.047	0.000	0.054
sex	-1.9158	0.335	-5.712	0.000	-2.573	-1.258
cp	0.8560	0.128	6.698	0.000	0.606	1.107
trestbps	-0.0114	0.007	-1.613	0.107	-0.025	0.002
chol	-0.0124	0.003	-3.859	0.000	-0.019	-0.006
fbs	-0.3403	0.373	-0.912	0.362	-1.072	0.391
restecg	0.3310	0.247	1.342	0.180	-0.153	0.814
thalach	0.0390	0.006	6.062	0.000	0.026	0.052
exang	-0.6096	0.285	-2.142	0.032	-1.167	-0.052
oldpeak	-0.7124	0.154	-4.629	0.000	-1.014	-0.411
slope	0.6537	0.242	2.697	0.007	0.179	1.129
ca	-0.7788	0.131	-5.936	0.000	-1.036	-0.522
thal	-0.8299	0.200	-4.142	0.000	-1.223	-0.437

Dep. Variable:	target
Method:	dydx
At:	overall

	dy/dx	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
age	0.0029	0.001	2.007	0.045	6.93e-05	0.006
sex	-0.2070	0.033	-6.267	0.000	-0.272	-0.142
cp	0.0925	0.012	7.823	0.000	0.069	0.116
trestbps	-0.0012	0.001	-1.621	0.105	-0.003	0.000
chol	-0.0013	0.000	-4.038	0.000	-0.002	-0.001
fbs	-0.0368	0.040	-0.914	0.361	-0.116	0.042
restecg	0.0358	0.027	1.348	0.178	-0.016	0.088
thalach	0.0042	0.001	6.782	0.000	0.003	0.005
exang	-0.0659	0.030	-2.172	0.030	-0.125	-0.006
oldpeak	-0.0770	0.015	-5.003	0.000	-0.107	-0.047
slope	0.0706	0.026	2.755	0.006	0.020	0.121
ca	-0.0842	0.013	-6.708	0.000	-0.109	-0.060
thal	-0.0897	0.020	-4.385	0.000	-0.130	-0.050

References

Gallis, J. A., & Turner, E. L. (2019, November). Relative measures of association for binary outcomes: Challenges and recommendations for the global health researcher. *Ann Glob Health*, *85*(1), 137. Retrieved from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6873895/`

Lapp, D. (2019, Jun). *Heart disease dataset.* Retrieved from `https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset`

Norton, E. C., Dowd, B. E., & Maciejewski, M. L. (2019). Marginal effects—quantifying the effect of changes in risk factors in logistic regression models. In E. H. Livingston & R. J. Lewis (Eds.), *Jama guide to statistics and methods.* New York, NY: McGraw-Hill Education. Retrieved from `jamaevidence.mhmedical.com/content.aspx?aid=1184195016`

Telke, S. E., & Eberly, L. E. (2011, September). Statistical hypothesis testing: Associating patient characteristics with a prevalent or incident condition—relative risk, odds ratio, and logistic regression. *J. Wound Ostomy Continence Nurs.*, *38*(5), 496–500. Retrieved from `https://oce-ovid-com.proxy1.library.jhu.edu/article/00152192-201109000-00006/HTML`

skimpy summary

Data Summary

dataframe	Values
Number of rows	1025
Number of columns	14

Data Types

Column Type	Count
int32	13
float64	1

number

column_name	NA	NA %	mean	sd	p0	p25	p75	p100	hist
age	0	0	54	9.1	29	48	61	77	
sex	0	0	0.7	0.46	0	0	1	1	
cp	0	0	0.94	1	0	0	2	3	
trestbps	0	0	130	18	94	120	140	200	
chol	0	0	250	52	130	210	280	560	
fbs	0	0	0.15	0.36	0	0	0	1	
restecg	0	0	0.53	0.53	0	0	1	2	
thalach	0	0	150	23	71	130	170	200	
exang	0	0	0.34	0.47	0	0	1	1	
oldpeak	0	0	1.1	1.2	0	0	1.8	6.2	
slope	0	0	1.4	0.62	0	1	2	2	
ca	0	0	0.75	1	0	0	1	4	
thal	0	0	2.3	0.62	0	2	3	3	
target	0	0	0.51	0.5	0	0	1	1	

End

Figure 1. Summary of Variables

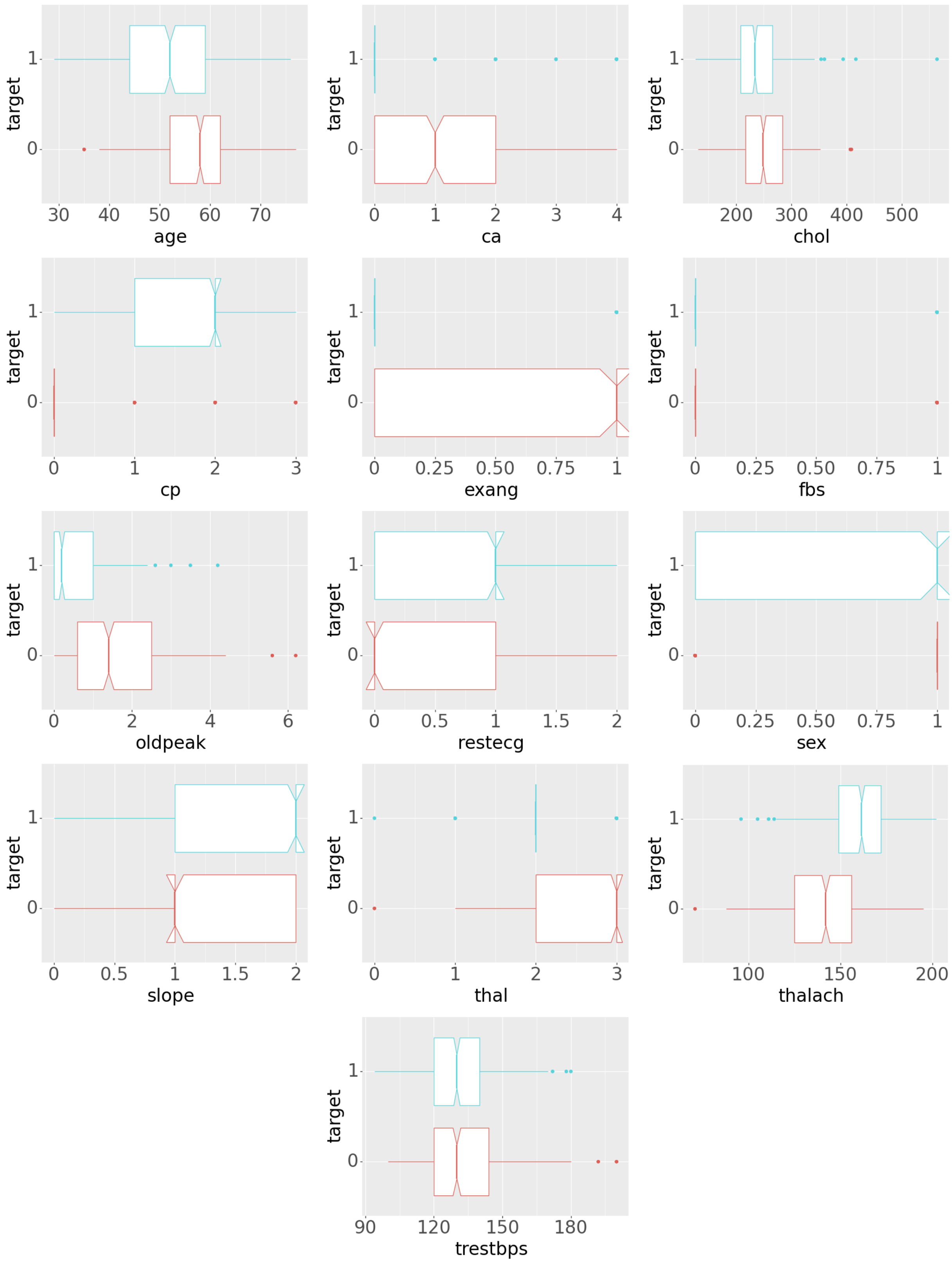

Figure 2. Distributions of Feature Variables

Appendix

Code

1. data_summary.py

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from skimpy import skim


# read data
df = pd.read_csv('data/heart.csv')
skim(df)

X = df.drop('target', axis=1)
y = df['target']

corr = df.corr()


def get_lower_tri_heatmap(data, output="src/plots/correlation.png"):
    mask = np.zeros_like(data, dtype=np.bool)
    mask[np.triu_indices_from(mask)] = True

    # Want diagonal elements as well
    mask[np.diag_indices_from(mask)] = False

    # Set up the matplotlib figure
    f, ax = plt.subplots(figsize=(11, 9))

    # Generate a custom diverging colormap
    cmap = sns.diverging_palette(220, 10, as_cmap=True)

    # Draw the heatmap with the mask and correct aspect ratio
    sns_plot = sns.heatmap(data, mask=mask, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5})
    # save to file
    fig = sns_plot.get_figure()
    fig.savefig(output)


if __name__ == '__main__':
    import plotnine as gg

    # plot all sub-distributions
    plots = []
    for col in X:
        plot = gg.ggplot(df) +\
               gg.geom_boxplot(gg.aes(y=col, x='factor(target)', color='factor(target)'), notch=True) +\
               gg.xlab('target') +\
```

```
                gg.coord_flip() +\
                gg.scale_color_discrete(guide=False) +\
                gg.theme(text=gg.element_text(size=24))
        plot.save(f'src/plots/target-{col}.png', 'png')

    # corr plot
    get_lower_tri_heatmap(corr)
```

2. associated_risk.py

```
import pandas as pd
from scipy.stats.contingency import relative_risk

from data_summary import X, y

# individual cross-tabs
age = pd.crosstab((X['age'] > 54).astype(int), y)
sex = pd.crosstab(X['sex'], y)
cp = pd.crosstab(X['cp'].isin([1, 2]).astype(int), y)
trestbps = pd.crosstab((X['trestbps'] > 130).astype(int), y)
chol = pd.crosstab((X['chol'] > 250).astype(int), y)
fbs = pd.crosstab(X['fbs'], y)
restecg = pd.crosstab(X['restecg'].isin([0, 1]).astype(int), y)
thalach = pd.crosstab((X['thalach'] > 150).astype(int), y)
exang = pd.crosstab(X['exang'], y)
oldpeak = pd.crosstab((X['oldpeak'] > 1.1).astype(int), y)
slope = pd.crosstab(X['slope'].isin([1, 2]).astype(int), y)
ca = pd.crosstab(X['ca'].isin([0, 1, 2]).astype(int), y)
thal = pd.crosstab(X['thal'].isin([2, 3]).astype(int), y)

crosstabs = [age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal]

# calculate ARD and RR
ard = lambda de, du, he, hu: round((de / (de + he)) - (du / (du + hu)), 4)
# rr = lambda de, du, he, hu: round((de / (de + he)) / (du / (du + hu)), 4)
risk_diff = dict()
risk_ratio = dict()
ci = dict()

for df in crosstabs:
    De = df.iloc[1, 1]
    Du = df.iloc[0, 1]
    He = df.iloc[1, 0]
    Hu = df.iloc[0, 0]

    var = df.index.name
    risk_diff[var] = ard(De, Du, He, Hu)
    rr = relative_risk(De, De + He, Du, Du + Hu)
    risk_ratio[var] = round(rr.relative_risk, 4)
    ci[var] = rr.confidence_interval(confidence_level=0.95)

risk = pd.DataFrame.from_records([risk_diff, risk_ratio, ci],
                                 index=['Risk Difference', 'Risk Ratio', 'RR 95% Confidence Interval'])
risk = risk.transpose()
risk['RR 95% Confidence Interval'] = risk['RR 95% Confidence Interval'].apply(lambda row: tuple([round(i, 4) for i in row]))
# risk['Correlation'] = X.corrwith(y)
# risk = risk[['Correlation', 'Risk Difference', 'Risk Ratio']]
risk.set_axis(['age', 'sex', 'chest pain     {1, 2}', 'resting blood
```

```
pressure > 130', 'serum cholestoral > 250 ml/dl',
              'fasting blood sugar > 120 mg/dl', 'resting electrocardiographic results     {0, 1}',
              'maximum heart rate achieved > 150', 'exercise induced angina', 'oldpeak', 'slope',
              'major vessels colored     {0, 1, 2}', 'thal     {2, 3}'], inplace=True)
risk.to_latex('src/risk-tables.tex', column_format='lrrr')
```

3. models.py

```python
import numpy as np
import pandas as pd
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, accuracy_score

from data_summary import X, y

# train/test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4,
    random_state=42)

# logistic regression
log_reg = sm.Logit(y_train, X_train).fit()
with open('src/logistic-regression.tex', 'w') as f:
    f.write(log_reg.summary().as_latex())

ame = log_reg.get_margeff(at='overall', method='dydx')
with open('src/marginal-effects.tex', 'w') as f:
    f.write(ame.summary().as_latex())

odds = np.exp(log_reg.params)
print(odds)

print(log_reg.summary())
print(ame.summary())
```